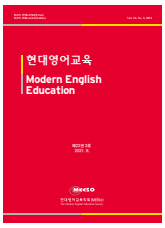
Modern English Education

https://doi.org/10.18095/meeso.2021.22.4.12



Development of a Multiple-Choice Writing Examination Validated with a Cognitive Diagnosis Model*

Sookyung Cho**
Hankuk University of Foreign Studies

Chanho Park
Keimyung University

ARTICLE INFO

Received: 30 September 2021
Revised: 28 October 2021
Accepted: 05 November 2021

Examples in: English
Applicable Languages: English
Applicable Levels: Elementary/ Secondary/Tertiary

KEYWORDS

multiple-choice exam/CDM/ second language writing/ validity/reliability
선다형 검사/인지진단모형/ 제2언어 쓰기/타당도/신뢰도

ABSTRACT

Cho, Sookyung, & Park, Chanho. (2021). Development of a multiple-choice writing examination validated with a cognitive diagnosis model. *Modern English Education*, *22*(4), 12-23.

The aim of this study was to explore the possibility of diagnosing second language (L2) learners' writing skills through a multiple-choice examination by applying a CDM (cognitive diagnosis model). L2 learners' writing abilities have mostly been assessed using direct methods such as essay tests. However, this method not only requires a large amount of human raters' labor, but also faces several issues such as intra-rater and inter-rater reliabilities. Essay examinations are also hard to validate because they are limited in checking the internal structure (i.e., conducting factor analysis). To compensate for these limitations, this study developed a multiple-choice examination and validated it with factor analysis and a CDM. The examination consisted of 20 items on key L2 writing components: content, organization, grammar, vocabulary, and mechanics. It was conducted on 109 college students. The CDM analysis revealed that the test was valid based on response processes. It provided substantial information about each test-taker's various skills. These results imply that L2 learners' writing abilities can be assessed in an indirect way and that the test can be conducted with a large body of students to provide useful information regarding their writing abilities at the same time.

I. INTRODUCTION

Although second language (L2) learners' writing abilities are mostly assessed with a direct method these days, in the mid-1900's, in particular, in the United States where there was a strong tradition of employing a psychometric approach in language tests, writing skills were assessed in an indirect way, like a multiple-choice

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020S1A5A2A01047467) and by Hankuk University of Foreign Studies Research Fund of 2021.
**First author: Sookyung Cho, Corresponding author: Chanho Park

Sookyung Cho (Professor)
Department of English Linguistics and Language Technology, Hankuk University of Foreign Studies, 107 Imun-ro, Dongdaemun-gu, Seoul, 02450, Korea
Email: sookyungcho@hufs.ac.kr / ISNI: 0000 0004 6109 256X

Chanho Park (Professor)
Department of Education, Keimyung University, 1095 Dalgubeol-daero, Dalseo-gu, Daegu, 42601, Korea
Email: cpark@kmu.ac.kr / ISNI: 0000 0004 6445 1724

test. For instance, Huddleston (1954) once showed that a multiple-choice test, consisting of items on punctuation, idiomatic expressions, grammar, and sentence structure, is highly correlated with the teachers' ratings of student writing. However, since the middle of the 20th century, writing scholars have begun disfavouring indirect assessments due to concerns regarding their construct validity (Eckes et al., 2016). Foremost, indirect assessments were criticized because of their focus on writing as a form. Since they are more likely to test writing skills as a composition of discrete skills—e.g., punctuation, grammar, expressions, or sentence structures, these indirect assessments were blamed for not being able to assess writing as a creative and communicative process.

Secondly, these indirect ways were reported not likely to have a positive impact on classroom instruction. Their numeric outcomes—e.g., total scores of test results, correlations with other external criteria such as washback effects or raters' judgment—were not necessarily considered helpful for teachers to tailor their instruction to each student's needs and appropriate levels. Recently, as formative assessment became popular, both writing scholars and teachers have become interested in using assessment not just as a tool of assessing students' writing skills, but as a diagnostic tool that informs the teacher of their students' strengths and weaknesses (Eckes et al., 2016). In this surge of interest in formative assessment, results of a multiple-choice exam have been reported insufficient to provide teachers with sufficient information about their students' writing skills.

Although multiple-choice exams may not directly assess writing abilities, they can be useful in classroom instruction since they are easier to apply cognitive diagnosis models (CDMs) to. CDMs refer to psychometric models that provide a profile of a learner's mastery of multiple attributes, and thus enable both teachers and test administrators to gain detailed information regarding the test-takers' knowledge, skills, and abilities (Rupp et al., 2010; von Davier & Y. S. Lee, 2019). Thus, multiple-choice writing exams used in conjunction with CDMs can provide teachers and learners of English with detailed diagnostic feedback in a large class taught even by novice teachers, and feedback obtained by applying a CDM can be used as a supplemental tool in writing classes (Manning & S. Cho, 2019). Therefore, this study aims to develop a multiple-choice exam that can be useful when teaching writing to L2 learners.

A developed exam can be validated by showing its evidence based on test content, response processes, internal structure, relations to other variables, and/or consequences of testing (American Educational Research Association et al., 2014). Thus, the multiple-choice exam is first validated for its internal structure by using factor analysis. CDMs cannot only be used to obtain diganostic feedback for the students, but also provide validity evidence based on response processes since a CDM can demonstrate what cognitive attributes the examinees are using to answer the items.

II. LITERATURE REVIEW

1. Methods of Assessing Second Language Writing Skills

Up until now, second language writing skills have been mostly assessed through a direct method because of the concerns regarding the construct validity of an indirect method, such as a multiple-choice exam. Scholars criticized that indirect methods did not assess writing as a creative and communicative process because they usually approached writing as a composite of discrete skills—e.g., punctuation, grammar, expressions, or sentence structures (Behizadeh & Engelhard, 2011). Additionally, indirect ways were criticized for not bringing a positive impact to classroom instruction (Bejar, 2012; Lumley, 2005; Wall, 2012). Their numeric outcomes—e.g., total scores of test results, correlations with other external criteria such as washback effects or raters' judgment—are not necessarily helpful for teachers to tailor their instruction to each student's needs and appropriate levels.

Due to these limitations, multiple-choice formats of assessing writing skills have been disfavored in the mid-20th century and instead interest in direct methods has surged. However, although essay exams elicit the test-takers' direct responses, that is, their writing skills, ironically, the process of assessing test-takers' performance cannot help but be at least somewhat indirect (Eckes et al., 2016). It involves the human raters' complex procedure of reading, evaluating, and scoring, which leads to unwanted rater variabilities: some raters are rather harsh while others are more lenient; raters provide similar scores on distinct criteria; or raters score essay based on their biases.

In order to solve these problems, either rater training or compensatory measurement models have been suggested and attempted (Bejar, 2012; Eckes et al., 2016; Lumley, 2005); however, these solutions have their own limitations as well. For example, the goal of rater training is to render the differences among raters negligible so that these rater variabilities would not contribute to variance irrelevant to the construct validity, but unfortunately, many studies have shown that rater training failed to accomplish this goal (Knoch, 2011; Lumley & McNamara, 1995; Weigle, 1998, 1999). For example, Weigle (1998) found that training did not reduce interrater reliability, in particular, in terms of rater harshness while intrarater reliability was increased. Additionally, Lumley and McNamara (1995) found that even this training effect did not last long, so they recommended regular training before each administration of a performance-based test. To make intrarater reliability more consistent, several measurement models have been applied, such as the generalizability theory (G-theory) or many-facet Rasch measurement (MFRM). G-theory recognizes multiple sources of measurement error and tries to make adjustments depending on their magnitude; on the other hand, the MFRM measures rater harshness and corrects test outcomes based on it. Even with such methodological modifications and adjustments attempted in the analysis

of direct writing assessment, it still has been pointed out that an essay exam, the direct method of assessing writing skills, cannot help but be indirect in its evaluation procedure (Eckes et al., 2016). The application of CDMs to a multiple-choice exam may not only solve these limitations of the direct method, but it may also help to overcome the aforementioned shortcomings of the multiple-choice exam.

2. Use of CDMs in Second Language Assessment

CDMs refer to "a family of psychometric models designed to provide categorical classifications for multiple latent attributes" (Paulsen & Valdivia, 2021). They provide a profile of a learner's mastery of multiple skills attributes, based on the analysis of a Q-matrix that indicates the appropriate attributes that are needed to answer a specific item. This detailed information regarding the test-takers' knowledge, skills, and abilities enables teachers to adjust their instruction and feedback in accordance with the students' levels and abilities (Rupp et al., 2010; von Davier & Y. S. Lee, 2019). Because of this advantage over a traditional format of tests that merely provide scores, CDMs have been used and studied in depth by language experts and scholars. However, the pioneering studies that first applied CDMs to second language proficiency assessment were mostly focused on assessing the learners' receptive skills, such as listening, reading, or grammar (C. Park & S. Cho, 2011, 2017; Sawaki et al., 2009). Because CDMs are based on dichotomous criteria, that is, the mastery or non-mastery of a certain skill or attribute, they have often been applied to these receptive skills of which the mastery can be easily decided upon from the results of multiple-choice exams; on the other hand, productive skills, including speaking or writing abilities, are likely to be tested in a direct way, and thus is not easy to discern whether a student mastered a certain attribute or not.

There have been several attempts to apply CDMs to the assessment of second language writing skills (Y. H. Kim, 2011; C. Park & S. Cho, 2020; Xie, 2016), but these procedures are limited in that they depend on human raters' labor and their subjective judgement. For example, Y. H. Kim (2011) extracted 35 attributes that are considered critical to second language writing skills with the help of a group of writing experts and then asked ten ESL teachers to evaluate 480 essays based on these 35 attributes. Similarly, in the study that applied a CDM to essays written by 472 Hong Kong college students, Xie (2016) also asked 10 teachers to read each essay and judge whether each writer mastered the same 35 attributes that were investigated in Y. H. Kim's study. Even these few cases clearly indicate that CDMs are not easily applicable to analyze essays of a large number of students because the procedure of deciding on each student's mastery of key attributes is as time-consuming and burdensome as those of assessing and scoring student essays. Also, if multiple raters are required to apply a CDM, problems such as inter-rater and intra-rater reliability may occur because the results are heavily dependent on the rater's subjective judgement.

However, the development of a multiple-choice test that has both reasonable levels of construct validity and reliability could ease these difficulties: in a multiple-choice test, it is simpler to decide whether a test-taker mastered a specific skill or not by looking at whether he or she answered the item correctly. As a result, the test can be applied to a large body of students and at the same time, a CDM can be applied to produce a similar level of information about how to teach each student as a direct method of writing assessment, such as an essay exam, would. For this purpose, this study aims to develop a multiple-choice exam that can assess an L2 learner's writing ability and then apply a CDM to analytically interpret these results. In this way, this study not only has the strengths of a quantitative measurement, in terms of reliability and validity, but also provides test administrators, teachers and learners with as rich and useful information as an essay exam does regarding test-takers' writing skills.

There are many types of CDMs, some of which have been applied to second language assessment. The most popular model is the deterministic inputs, noisy and gate (DINA) model (de la Torre, 2009). The formulation of the DINA model starts with defining η_{ii}, a latent variable showing whether an examinee i (=1, 2, …, I) mastered the attributes necessary to correctly answer item j (=1, 2, …, J) like the following Eq. 1.

$$\eta_{ii} = \Pi_{k=1}^{K} (\alpha_{ik})^{q_{jk}}. \quad (1)$$

In (1), α_{ik} is a binary indicator variable of whether examinee i mastered attribute k (=1, 2, …, K), and q_{jk} is an element of the Q-matrix, whose dimension is $J \times K$ and which is necessary for successfully answering the items.

In order to formulate a successful answer to item j, additional noise parameters are defined as follows.

$$g_j = P(x_{ij}=1 \mid \eta_{ij}=0) \quad (2)$$
$$s_j = P(x_{ij}=0 \mid \eta_{ij}=1), \quad (3)$$

where g_j and s_j denote the probability of correctly answering the item x_{ij}, when the examinee did not master all the required attributes and the probability of incorrectly answering even though the examinee mastered all the required attributes, respectively. Thus g_j and s_j are called the guessing parameter and the slipping parameter.

Finally, the DINA model is defined like Eq. 4 in the following.

$$P(x_{ik}=1 \mid \eta_{ij}) = g_j^{1-\eta_{ij}} (1-s_j)^{\eta_{ij}}, \quad (4)$$

Simply put, it is defined as the product of the probability of correctly guessing the answer of an item when the examinee cannot find the correct answer and the probability of not slipping when the examinee has the correct answer.

A generalized DINA (G-DINA) model relaxes some assumptions of the DINA model. The DINA model assumes that the examinees who did not master any of the required attributes are equally incompetent; i.e., they have the same

probability of success for the item. However, examinees can have different probabilities of success depending on which attributes the examinees did not master. As a general CDM, the G-DINA has other CDMs as its special cases. Those models are the DINA, additive CDM (ACDM; de la Torre, 2011), reduced reparameterized unified (RRUM; Hartz, 2002), etc. Explaining the models in detail is beyond the scope of this study, and interested readers can refer to de la Torre (2011) for the details and the relationship of these CDMs.

III. METHOD

1. Participants

The participants of this study are 109 college students, ranging from freshmen to seniors, who were enrolled in various English writing-related courses—e.g., Critical Writing, Advanced English Writing, Essay Reading and Writing, or Research Writing—at a university located in Seoul, Korea. They voluntarily participated in this study. There were 58 females and 51 males, whose ages ranged from 18 to 30. In their first writing classes, they were recruited to participate in this study with the help of their writing instructors, and they took an online multiple-choice writing exam for about 30 minutes during their class time.

2. Instrument

A multiple-choice writing exam was developed for this study. It consists of 20 items on five different areas—content (CON), organization (ORG), grammar (GRM), vocabulary (VOC), and mechanics (MCH), which have been confirmed to be critical to second language writing in several CDM studies (Y. H. Kim, 2011; C. Park & S. Cho, 2020; Xie, 2016). Three graduate assistants participated in the development of the multiple-choice exam; all of them not only have had training in teaching L2 writing, but also had experiences of teaching English writing to Korean students. In the first round, they were given guidelines on what to test on this exam and were asked to develop several multiple-choice items on their own. In the second round, the authors of this study examined the qualities and appropriateness of these items and selected 20 items that were considered fit for the purpose of this study. In the third round, the test of 20 items was conducted on 20 students of one of the writing classes mentioned above as a pilot study and was thereby confirmed to be appropriate to assess a Korean college student's English writing ability (See the Appendix for the detail).

To apply CDMs, a Q-matrix, a binary incidence matrix for the relationship between the items and the item attributes, is necessary. The attributes were taken from previous studies (e.g., C. Park & S. Cho, 2020). When developing the multiple-choice items, the Q-matrix in Table 1 was constructed simultaneously. The researchers of this study independently checked the validity of the Q-matrix, which was also confirmed by the graduate assistants.

TABLE 1
Q-matrix Used in This Study

Item	Attributes				
	CON	ORG	GRM	VOC	MCH
1	1	1	0	0	0
2	1	1	0	0	0
3	0	0	1	1	0
4	0	0	1	1	0
5	1	1	0	0	0
6	0	0	1	0	0
7	0	1	0	1	0
8	1	1	0	0	0
9	1	1	0	0	0
10	0	0	1	0	1
11	0	1	0	1	0
12	1	1	0	0	0
13	1	1	0	0	0
14	0	1	1	1	0
15	0	0	1	0	0
16	0	1	0	1	0
17	0	1	0	1	0
18	1	1	0	0	0
19	1	1	0	0	0
20	1	1	0	0	0

Note. CON = content, ORG = organization, GRM = grammar, VOC = vocabulary, MCH = mechanics

3. Data Analysis

The steps of data analysis are as follows. First, Cronbach's α coefficient was computed for the reliability analysis, and factor analysis was conducted for the validity analysis. Since the variables are dichotomous items, nonlinear factor analysis based on tetrachoric correlation coefficients was carried out with an oblique rotation. Second, item-level statistical analysis results such as difficulty (*p*-value) and discrimination (item scores correlated with the total score subtracting the item score) were obtained. Then, CDMs were applied for the last analysis. CDMs such as the DINA, G-DINA, ACDM, and RRUM were compared for model fit, and the best-fitting model was chosen and applied to the data. Mplus 8.3 (Muthén & Muthén, 1998-2017) was used for the factor analysis, and all the other analyses were conducted using R 4.1.1 (R Core Team, 2021). In particular, R packages, GDINA and CDM, were applied for CDM analyses (Ma & de la Torre, 2020; Robitzsch et al., 2020).

IV. RESULTS

1. Basic Statistics

First, sum scores were computed for the dichotomously scored data, and Table 2 shows their descriptive statistics. Out of the 20 items in total, no examinees scored higher than 18 or lower than 2; the average score was 8.65 with the standard deviation of 3.50 and the median was 9. Figure 1 shows the sum score's histogram with a hypothetical normal curve in red. The distribution of sum scores looks close to a normal distribution except that there are quite a few low-achieving students.

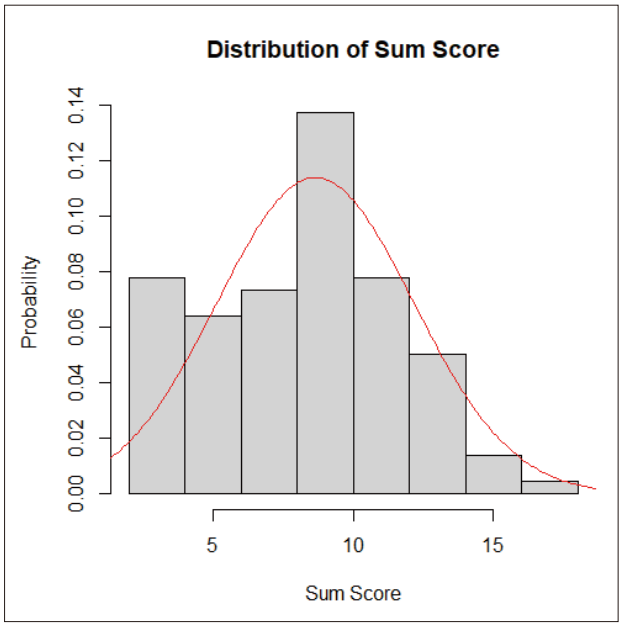


FIGURE 1 Distribution of Sum Score

TABLE 2
Descriptive Statistics of Sum Scores

M	SD	Median	Min	Max
8.65	3.50	9	2	18

2. Reliability and Validity

The test's Cronbach α coefficient as a reliability estimate was .673. It was acceptable because this exam was not a high-stakes test but a classroom assessment for a formative purpose, the sample size was not large, and Cronbach α is not a reliability estimate itself but a lower bound to reliability (McDonald, 1999).

To obtain validity evidence based on the internal structure of the test, nonlinear factor analysis was conducted with an oblique rotation (Mplus uses GEOMIN rotation method as a default). Table 3 shows the comparison between a 1-factor model and a 2-factor model. Using .05 as a nominal Type 1 error rate, the 2-factor model retains the null hypothesis of model fit, and the chi-square difference test favors the 2-factor model as well, which may have caused the low Cronbach α coefficient.

TABLE 3
Comparison of Factor Analysis Models

Model	χ^2	*df*	*p*	$\Delta\chi^2$	Δdf	*P*
1-factor	202.636	170	.0442	35.448	19	.0123
2-factor	164.353	151	.2162			

Table 4 shows the factor loadings as well as the difficulty (*p*-value) and discrimination (adjusted correlation with the sum score). Items 1 and 13 show negative discrimination indices, and items 12 and 20 have discriminations under .10. As was shown in the sum score distribution, this test was slightly difficult for the examinees with items 2 and 8 as the easiest and hardest items, respectively. Factor loadings (only the ones whose absolute values are larger than .20 are shown) show that this test is possibly testing two constructs. It is common that a reading comprehension examination is multidimensional; that is, because reading items usually share a common passage, they tap multiple constructs, such as an ability to solve a particular item and another ability to understand the passage. The writing test in this study has one common reading passage for all items, which might have caused the multidimensionality of the test. The correlation coefficient between the two factors was .256. The analysis of the items positioned in Factor 1 implies that the first factor relates to the test-takers' background knowledge of the topic dealt with in the test passage, that is, food processing, whereas the second factor is related to coherence between the sentences, how smoothly the transition from one sentence to the other one. The results of factor analysis indicate that there may exist these two factors (test-takers' knowledge of topic and coherence) as sub-categories under the five composites used in the Q-matrix for this study—content, organization, vocabulary, grammar, and mechanics—in particular, content and organization.

TABLE 4
Factor Loadings with Item Statistics

Item	Difficulty	Discrimination	Rotated factor loading	
			Factor 1	Factor 2
1	0.303	-0.014	0.536*	-0.291
2	0.780	0.431		0.776*
3	0.211	0.261	0.880*	
4	0.404	0.196	-0.276	0.549*
5	0.404	0.167	0.289	
6	0.358	0.471	0.224	0.595*
7	0.413	0.196	0.222	
8	0.156	0.213		0.415*
9	0.459	0.298		0.612*
10	0.477	0.510	0.258	0.637*
11	0.431	0.129		0.331*
12	0.211	0.032	0.357*	
13	0.275	-0.072	-0.279	
14	0.468	0.328	0.351*	0.302
15	0.385	0.260	0.397*	0.239
16	0.688	0.495	0.254	0.611*
17	0.697	0.399	0.558*	0.294
18	0.523	0.281		0.523*
19	0.615	0.335		0.525*
20	0.394	0.099	-0.236	0.325*

* $p < .05$.

3. CDM Analysis

For the CDM analysis, four models (DINA, G-DINA, ACDM, and RRUM) were compared first, and the results are shown in Table 5. Akaike's Information Criterion (AIC; Akaike, 1973), the Bayesian Information Criterion (BIC; Schwarz, 1978), Bozdogan's consistent AIC (CAIC; Bozdogan, 1987) were used to compare the models. Although AIC suggested that G-DINA was the best model fit, AIC is known to erroneously favor a model with too many parameters when the sample size is small (Giraud, 2015). In this study, the sample size was 109 and G-DINA has 96 parameters—almost double that of DINA, so instead of the GINA, the DINA model was used. Both CAIC and BIC favored DINA as the best model. The DINA model being the best fitting CDM to the data shows evidence that the examinees are applying the five attributes (content, organization, grammar, vocabulary, and mechanics) to answer the examinees, which shows validity evidence based on response processes.

Table 6 shows the item parameters (*g* for guessing and s for slipping), their standard errors as well as RMSEA (root mean squared error of approximation) and IDI (item discrimination index), which is simply computed as 1 - *g* - *s*. Note that items with low RMSEA and high IDI are considered 'good' items. The slipping parameters are generally higher than the guessing parameters, which shows the possibility that test-takers needed other attributes than the ones specified in this model to successfully answer the items. Figure 2 is a visual representation of the item parameters showing guessing and non-slipping (= 1 - *s*) parameters in Table 6.

TABLE 5
Comparison of CDM Results

Model	Deviance	# of Parameters	AIC	CAIC	BIC
DINA	2585.532	56	2697.532	**2904.248**	**2848.248**
G-DINA	2493.185	96	**2685.185**	3039.554	2943.554
ACDM	2540.020	75	2690.020	2966.871	2891.871
RRUM	2544.193	75	2694.193	2971.044	2896.044

Note. Bold type represents the lowest number.

TABLE 6
Item Statistics by the DINA Model

Item	*g*	SE(*g*)	*s*	SE(*s*)	RMSEA	IDI
1	0.294	0.084	0.693	0.056	0.084	0.013
2	0.353	0.091	0.041	0.026	0.079	0.606
3	0.062	0.043	0.718	0.053	0.109	0.220
4	0.227	0.081	0.512	0.059	0.078	0.261
5	0.210	0.076	0.515	0.060	0.067	0.275
6	0.047	0.042	0.511	0.058	0.092	0.442
7	0.246	0.091	0.523	0.058	0.095	0.231
8	0.109	0.058	0.824	0.045	0.086	0.067
9	0.134	0.061	0.405	0.059	0.092	0.461
10	0.000	0.000	0.148	0.056	0.004	0.852
11	0.323	0.099	0.527	0.058	0.035	0.150
12	0.206	0.076	0.787	0.049	0.081	0.007
13	0.202	0.075	0.694	0.055	0.047	0.104
14	0.218	0.075	0.413	0.060	0.064	0.369
15	0.064	0.049	0.479	0.058	0.093	0.457
16	0.215	0.087	0.130	0.038	0.043	0.655
17	0.376	0.103	0.179	0.044	0.089	0.445
18	0.282	0.084	0.376	0.058	0.056	0.342
19	0.316	0.088	0.260	0.053	0.071	0.424
20	0.323	0.086	0.576	0.059	0.093	0.101

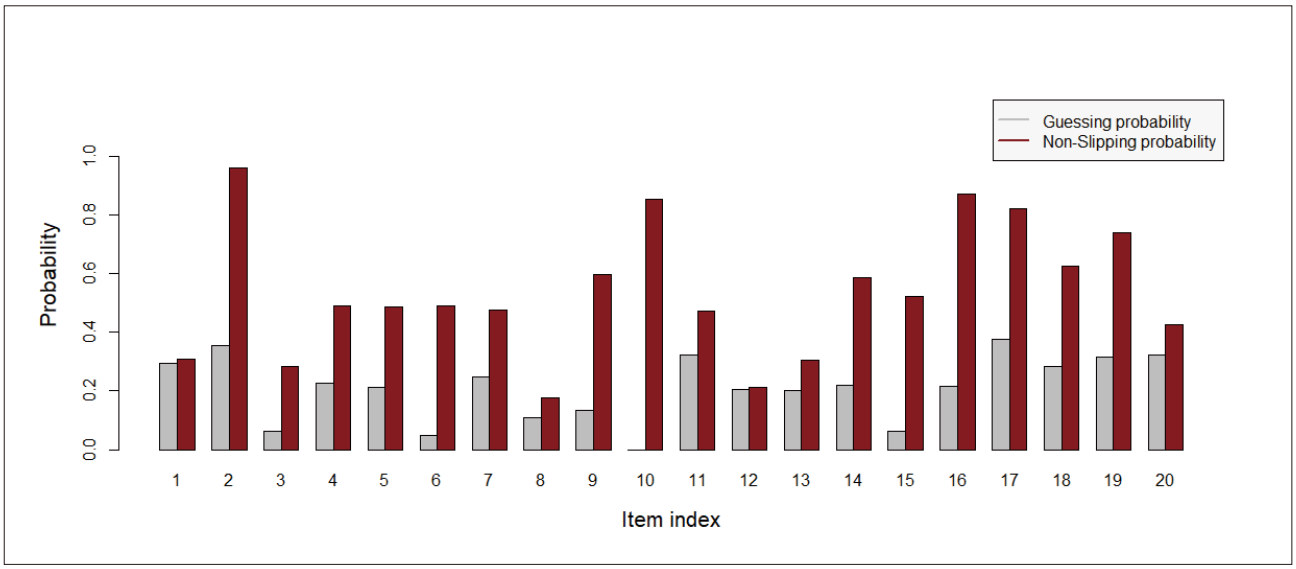


FIGURE 2 Guessing and Non-Slipping Parameters per Item

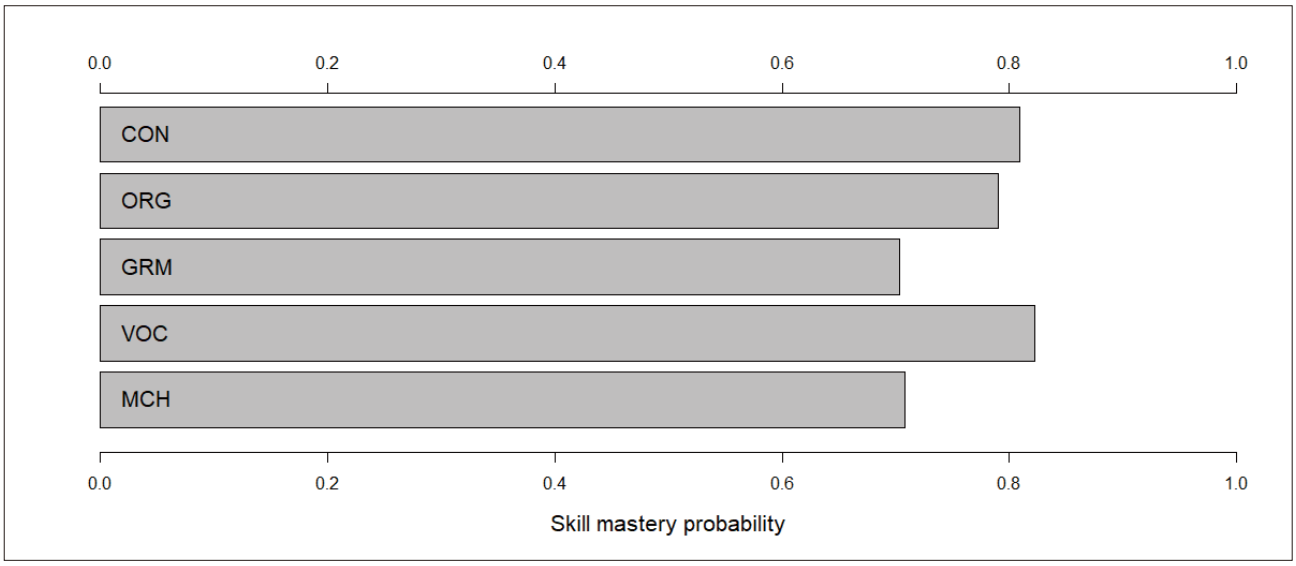


FIGURE 3 Posterior Probabilities for Skill Mastery

Finally, Figure 3 shows posterior probabilities for each skill (or attribute). Mastery of content and vocabulary were more common than that of grammar and mechanics. Also, the posterior probability for the skill pattern class was .523 for the class of 11111 (the class who mastered all the five attributes), followed by .122 for the class of 11110 (the class who did not master MCH only). The posterior probabilities for the other classes were small, but the classes whose probabilities were larger than .02 were 00000, 10000, 01000, 00010, 00001, 10010, 10001, 01001, 00011, 11010, 10011, 11011, and 01111. The attributes produce 32 (=2^5) possible latent classes, but only less than half of the classes have nonignorable class sizes. This information can be useful both for the students and the teachers; that is, students can be informed of what areas they have mastered or not while the teachers will be able to tailor their instruction and feedback to the students' mastery levels.

V. Discussion and Conclusion

A series of statistical analyses, such as Cronbach α coefficient, factor analysis, and the CDM, shows that the writing test developed in this study has appropriate levels of reliability and validity in terms of assessing L2 learners' writing skills. According to American Educational Research Association et al. (2014), assessment tools need validity evidence based on test content, internal structure, relations to other variables, response processes, and consequences of testing. Although the first three pieces of evidence have often been adopted in many testing studies in order to validate their assessment tools, response processes have hardly been obtained because it is hard to observe and investigate the test-takers' cognitive processes behind their responses in an exam. However, this study obtained validity evidence based on response processes through the CDM, and validity evidence based on internal structure

through factor analysis. The CDM is a model of understanding a student's mind and cognition since it shows us how examinees respond to the test items (Rupp et al., 2010). By adopting the CDM, this study could understand what skills and attributes examinees used to answer each item, what skills were the students most likely to have mastered, or what classes students could be classified into based on their mastered attributes and skills. Such information contributed to understanding their response processes while answering the questions in the multiple-choice writing exam developed in this study.

The development of a multiple-choice writing exam for the purpose of assessing L2 learners' writing skills has brought several advantages to this study. First, the use of the multiple-choice writing exam makes it possible to obtain the validation procedure through a series of statistical analyses, which has been mentioned above. Although essay exams are assumed to be a more appropriate tool for assessing L2 writing skills, it has been lamented that they cannot go through such rigorous and statistical validation procedures as conducted in this study (Ackerman & Smith, 1988; Eckes et al., 2016; Palmer & Devitt, 2007). Second, the use of the CDM helps to solve the problem of an indirect method of assessing writing skills, that is, its insufficiency to affect classroom instruction due to the lack of information regarding students' writing skills (Eckes et al., 2016). The CDM results provide teachers with fine-grained information regarding their students' current levels of L2 writing skills by showing them what key attributes their students have mastered so far. They can utilize this information in many aspects; for example, they can tailor their instruction and feedback to the students' levels or place the students of similar mastery patterns into different classes or groups. Lastly, the results of this study can be applied to a large body of students. Usually, essay exams require a group of well-trained raters in order to grade them, and thus, they become cumbersome to conduct and grade as their scale becomes larger. Moreover, it is almost impossible to give individual students detailed feedback when there are hundreds and thousands of students. However, the CDM analysis of the multiple-choice exam results as conducted in this study provides standardized quality feedback in a relatively quick and easy manner, compared to a direct essay exam.

Even so, this study has several limitations, which require further studies. First, the sample size is small. More reliable results could be obtained with more students. Second, the multiple-choice exam used in this study consists of 20 items on only one passage. As seen in the test results, one of the two factors that play important roles in students' answering the questions, relates to their background knowledge of the topic. If they were tested on several passages of various topics, it could have been investigated whether this factor works regardless of passage topics. Lastly, more detailed attributes might have helped better understand the students' L2 writing skills. In this study, content, organization, grammar, vocabulary, and mechanics were used as attributes to explain how students responded to the writing test items. These attributes are not general terms. That is, 'grammar' in this study does not mean grammar in general but grammar necessary to solve the items in the writing test. For this reason, more detailed attributes will help us better understand how students solve and respond to these items.

These limitations notwithstanding, this study is a successful approach that mitigates each disadvantage of a direct and indirect method of assessing L2 writing skills. That is, although a multiple-choice writing exam is easy to conduct and handle, it could not provide as rich information regarding a test-taker's writing skill as an essay exam could. By applying the CDM to multiple-choice test results, this study, however, opens the possibility of taking advantage of the strength of a multiple-choice test, handiness in conducting the exam, at the same time overcoming its disadvantage, the limited information about students. And thus, the findings of this study imply that multiple-choice exam results, like this study's, could help writing teachers to guide their instruction and successfully work as a formative assessment in a writing classroom.

REFERENCES

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing test. *Applied Psychological Measurement, 12*(2), 117-128.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267-281). Akademai Kiado.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Behizadeh, N., & Engelhard, G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing, 16*(3), 189-211.

Bejar, I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 16*(3), 189-211.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345-370.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115-130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179-199.

Eckes, T., Müller-Karabil, A., & Zimmerman, S. (2016). Assessing writing. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 147-164). De Gruyter Mouton.

Giraud, C. (2015). *Introduction to high-dimensional statis-*

tics. Boca Raton, FL: CRC.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* [Unpublished doctoral dissertation]. University of Illinois at Urbana-Champaign.

Huddleston, E. M. (1954). Measurement of writing ability at the college-entrance level: Objective vs. subjective testing techniques. *Journal of Experimental Education, 22*(3), 165-213.

Kim, Youn-Hee (2011). Diagnosing EAP writing ability using the reduced reparameterized unified model. *Language Testing, 28*(4), 509-541.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behaviour: A longitudinal study. *Language Testing, 28*(2), 179-200.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective.* Lang.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software, 93*(14), 1-26.

Manning, S. J., & Cho, Sookyung (2019). Engagement with automated writing feedback in mandated vs. voluntary conditions. *Modern English Education, 20*(4), 18-30.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Lawrence Erlbaum Associates Publishers.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Palmer, E. J., & Devitt, P. G., (2007). Assessment of higher order cognitive skills in undergraduate education: Modified essay or multiple choice questions? Research paper. *BMC Medical Education, 7*(49), 1-7.

Park, Chanho, & Cho, Sookyung (2011). Cognitive diagnostic assessment of English grammar for Korean EFL learners. *English Teaching, 66*(4), 101-117.

Park, Chanho, & Cho, Sookyung (2017). An exploratory analysis of compensatory cognitive diagnosis in EFL reading comprehension. *Korean Journal of Language and Linguistics, 17*(1), 85-104.

Park, Chanho, & Cho, Sookyung (2020). Cognitive diagnostic writing assessment for Korean learners of English. *Studies in Foreign Language Education, 34*(2), 1-24.

Paulsen, J., & Valdivia, D. S. (2021). Examining cognitive diagnostic modelling in classroom assessment conditions. *The Journal of Experimental Education,* DOI: 10.1080/00220973.2021.1891008

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Robitzsch, A., Kiefer, T., George, A. C., & Ünlü, A. (2020). CDM: *Cognitive diagnosis modeling*. R package version 7.5-15, https://CRAN.R-project.org/package=CDM.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications.* Guilford.

Sawaki, Y., Kim, Hae-Jin, & Gentile, C. (2009). Q-matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessment. *Language Assessment Quarterly, 6*(3), 190-209.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461-464.

von Davier, M., & Lee, Young-Sun (Eds.). (2019). *Handbook of diagnostic classification models: Models and model extensions, applications, software packages*. Springer.

Wall, D. (2012). Washback. In G. Fulcher & F. Davidson. (Eds.), *The Routledge handbook of language testing* (pp. 79-92). Routledge.

Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.

Weigle, S. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178.

Xie, Q. (2016). Diagnosing university students' academic writing in English: Is cognitive diagnostic modelling the way forward? *Educational Psychology, 37*(1), 26-47.

APPENDIX

Multiple-Choice Writing Exam

Read the passage and choose the answer to each question that most effectively improves the quality of writing in the passage and that makes the passage confirm to the conventions of standard written English. Each question includes a "NO CHANGE" option. Choose that option if you think the best choice is to leave the relevant portion of the passage as it is.

Whey to Go

[1] Greek yogurt—a strained form of cultured yogurt—has grown enormously in popularity in the United States since it was first introduced in the country in the late 1980s.

[2] From 2011 to 2012 alone, sales of Greek yogurt in the US increased by 50 percent. The resulting increase in Greek yogurt production has forced those involved in the business to address the detrimental effects **[3]** that the yogurt-making process may be having on the environment. (#1) Fortunately, farmers and others in the Greek yogurt business have found many methods of controlling and eliminating most environmental threats. (#2) Given these solutions as well as the many health benefits of the food, the advantages of Greek yogurt **[4]** outdo the potential drawbacks of its production. (#3)

(#4) The main environmental problem caused by the production of Greek yogurt is the creation of acid whey as a by-product. (#5) **[5]** Because it requires up to four times more milk to make than conventional yogurt does, Greek yogurt produces larger amounts of acid whey, which is difficult to **[6]** dispose of. (#6) **[7]** To address the problem of disposal, farmers have found a number of uses for acid whey. They can add it to livestock feed as a protein **[8]** supplement, and people can make their own Greek-style yogurt at home by straining regular yogurt. **[9]** If it is improperly introduced into the environment, acid-whey runoff **[10]** can pollute waterways, **[11]** eliminating the oxygen content of streams and rivers as it decomposes. Yogurt manufacturers, food scientists, and government officials are also working together to develop additional solutions for reusing whey. **[12]**

[13] Though these conservation methods can be costly and time-consuming, they are well **[14]** worth the effort. (#7) Nutritionists consider Greek yogurt to be a healthy food: it is an excellent source of calcium and protein, serves as a digestive aid, and contains few calories in its unsweetened low-and non-fat forms. (#8) Greek yogurt is slightly lower in sugar and carbohydrates than **[15]** conventional yogurt is. **[16]** Also, because it is more concentrated, Greek yogurt contains slightly more protein per serving, thereby helping people stay **[17]** satiated for longer periods of time. **[18]** These health benefits have prompted Greek yogurt's recent surge in popularity. **[19]** In fact, Greek yogurt can be found in an increasing number of products such as snack food and frozen desserts. Because consumers reap the nutritional benefits of Greek yogurt and support those who make and sell it, farmers and businesses should continue finding safe and effective methods of producing the food. (#9) **[20]**

1. The writer is considering elaborating on this introduction. Should the writer do this?

A) Yes, because it does not provide enough information about Greek yogurt.

B) Yes, because it fails to support the main argument of the passage.

C) No, because it provides enough information about Greek yogurt.

D) No, because it sets up the argument in the paragraph for the popularity of Greek yogurt.

2. For the sake of cohesion in this paragraph, this sentence should be placed

A) where it is now.

B) in (#1).

C) in (#2).

D) in (#3).

3.

A) NO CHANGE

B) that the yogurt-making process on the environment

C) that the yogurt-making process having on the environment

D) that the yogurt-making process could have on the environment

4.

A) NO CHANGE

B) outperform

C) outweigh

D) defeat

5. Which choice provides the most relevant detail?
A) NO CHANGE
B) Like other dairy products, Greek yogurt contains natural hormones, which can be harmful to people with hormone imbalances.
C) As it decomposes, acid whey removes oxygen from water, wreaking havoc on aquatic ecosystems.
D) The boom in Greek yogurt production has led to a threefold increase in New York alone of this toxic liquid between 2007 and 2013.

6.
A) NO CHANGE
B) dispose of it
C) dispose
D) disposal

7.
A) NO CHANGE
B) solve
C) undertake
D) engage in

8. Which choice provides the most relevant detail?
A) NO CHANGE
B) supplement and convert it into gas to use as fuel in electricity production.
C) supplement, while sweet whey is more desirable as a food additive for humans.
D) supplement, which provides an important element of their diet.

9. To make this paragraph most logical, this sentence should be placed
A) where it is now.
B) in (#4).
C) in (#5).
D) in (#6).

10.
A) NO CHANGE
B) can pollute waterways
C) could have polluted waterways,
D) have polluted waterway's

11.
A) NO CHANGE
B) eradicating
C) depleting
D) shortening

12. At this point, the writer is considering adding the following sentence.

> Therefore, it has been shown that acid whey can be reused in smoothies, mixed drinks and lacto-fermented soda, and it is found that people's hair can be softer and brighter if rinsing with acid whey.

Should the writer make this addition here?
A) Yes, it is necessary because it helps to show acid whey can be reused in many ways.
B) Yes, it is necessary because it suggests that only Greek yogurt is a product that can save farms.
C) No, it is unnecessary because it shifts the focus of acid whey from being reused on farms to being reused for people.
D) No, it is unnecessary because it is not related to the entire content of the passage.

13. The writer is considering deleting this sentence. Should the writer do this?
A) Yes, because it does not provide a transition from the previous paragraph.
B) Yes, because it fails to support the main argument of the passage as introduced in the first paragraph.
C) No, because it continues the explanation of how acid whey can be disposed of safely.
D) No, because it sets up the argument in the paragraph for the benefits of Greek yogurt.

14.
A) NO CHANGE
B) worthy
C) worthwhile
D) worthless

15.
A) NO CHANGE
B) conventional yogurt
C) conventional yogurt has
D) conventional yogurt is lower

16.
A) NO CHANGE
B) In other words,
C) Therefore,
D) For instance,

17.
A) NO CHANGE
B) satiating
C) complacent
D) sufficient

18. At this point, the writer is considering adding the following sentence.

> Eventually, Greek yogurt makes people save their money on food because they are not easily hungry.

Should the writer make this addition here?
A) I'm not sure whether this sentence is necessary or not.
B) Yes, because it explains the advantages of Greek yogurt more.
C) Yes, because it is a key point that gives readers new information.
D) No, because it is not connected to the sentence that follows and does not explain whether it is true or not.

19. To make this paragraph most logical, this sentence should be placed
A) where it is now.
B) in (#7)
C) in (#8).
D) in (#9).

20. The writer is considering elaborating on the current conclusion. Should the writer do this?
A) Yes, because it does not restate the writer's argument.
B) Yes, because it fails to provide the writers' final thoughts on Greek yogurt industry.
C) No, because it clearly shows and effectively supports the writer's argument.
D) No, because it concludes the passage with an appropriate prediction in the future.